الگوریتم بازیابی و رتبه‌بندی اطلاعات در موتور جستجوی گوگل

سعیده ابراهیمی[1]

چکیده

هدف اساسی این مقاله، تبیین نحوهٔ عمل موتور جستجوی گوگل در رتبه‌بندی اطلاعات بازیابی شده است و بدین منظور، الگوریتم (خوارزمی) موتور جستجوی گوگل را در بازیابی و رتبه‌بندی اطلاعات بررسی می‌کند. در بخش اول - که text matching نامیده می‌شود- شیوهٔ یافتن اطلاعات مرتبط با واژه‌های وارد شده درجعبهٔ جستجو، به‌طور خلاصه بیان می‌گردد و در بخش دوم - که بخش اصلی مقاله است و PageRank نام دارد - شیوهٔ رتبه‌بندی نتایج مرتبط بازیابی می‌شود و به‌تفصیل مورد بررسی قرار می‌گیرد. درادامه، الگوی موج‌سوار تصادفی، به‌عنوان تأییدی بر الگوریتم (خوارزمی) رتبه‌بندی گوگل تشریح می‌گردد. این مقاله در پایان، روش مشخص کردن رتبهٔ صفحات وب را از طریق نوار ابزار گوگل توصیف می‌کند.

کلیدواژه‌ها

موتور جستجوی گوگل، الگوریتم، PageRank، رتبه‌بندی اطلاعات، لینک، الگوی موج‌سوار تصادفی.

مقدمه

در طی چند سال اخیر، گوگل به پراستفاده‌ترین موتور جستجوی جهان تبدیل شده است. این مسئله، نه تنها به عملکرد و سهولت استفاده از این موتور برمی‌گردد، بلکه کیفیت نتایج جستجو در آن نیز تا اندازهٔ زیادی بر این مسئله سایه انداخته است.

گوگل، هم برای کاربران و هم برای مدیران وبگاه‌ها دارای ارزشی منحصربه‌فرد است. خزنده‌های موجود در موتور گوگل،

۱. دانشجوی دکتری کتابداری و اطلاع‌رسانی دانشگاه شهید چمران اهواز. ebrahimy_saedeh@yahoo.com

با سرعت قابل ملاحظه‌ای، وبگاه‌ها را نمایه و فهرست می‌کنند. اگر یک URL به‌تازگی ساخته شده باشد، درکمتر از دو هفته در دستگاه نمایه‌سازی گوگل، نمایه می‌شود. وبگاه‌های نمایه شده را نیز هرماه یک‌بار، با هدف درج تغییرات، از نو نمایه‌سازی می‌کند. این جریان نمایه‌سازی، بسیار سریع‌تر از موتورهای دیگر انجام می‌گیرد. (۳)

گوگل، روزانه به ۴۰ میلیون جستجو پاسخ می‌دهد (۱۱: ۱۰۸). سرعت بازیابی در موتورگوگل، کمتر از ۰/۵ ثانیه است. این موتور ادعا دارد که امکان جستجو به ۳۵ زبان مختلف را نیز فراهم کرده است. (۱)

نتایج جستجو در موتور گوگل از لحاظ میزان ارتباط به نسبت موتورهای دیگر تقریباً در بالاترین سطح قرار دارد. در تحقیقی که دیمتریوس و گوتلیب انجام دادند و موتور گوگل، یاهو و لایکوز را مقایسه کردند، در جستجو با کل واژگان، جستجوی مجاورتی و مکانی ساده و بولی، موتور گوگل از لحاظ میزان ارتباط، بالاترین امتیاز را کسب کرد. (۵: ۴۲)

موفقیت گوگل در بازیابی اطلاعات، از در اختیار داشتن بهترین الگوریتم جهت رتبه‌بندی نتایج بازیابی ناشی شده که بر ماهیت کاملاً آزادانهٔ وب استوار است.

رتبه‌بندی صفحات وب در موتورهای جستجو

رتبه‌بندی صفحات بازیابی شدهٔ وب در پاسخ به هر جستجو، عامل مهمی است که می‌تواند رضایت کاربران و در نهایت موفقیت موتور جستجو را تضمین کند. «با توجه به نتایج تحقیقات در این زمینه که نشان می‌دهد کاربران از چندین صفحهٔ بازیابی کردهٔ موتورهای جستجو، تنها تعداد محدودی از صفحات اولیه را مرور می‌کنند» (۴: ۳۹)، مسئلهٔ رتبه‌بندی نتایج، اهمیت دوچندان پیدا می‌کند.

سیاههٔ نتایج بازیابی شده، براساس معیارهای متنوع خاص هر موتور جستجو رتبه‌بندی می‌شود. از جملهٔ این معیارها، می‌توان به تعداد واژگان منطبق شده با واژگان جعبهٔ جستجو، مجاورت واژگان، محل قرار گرفتن واژگان در مدرک، بسامد واژگان (در هر مدرک و در کل پایگاه) و طول مدرک اشاره کرد (۴: ۳۹). کیم[2] (۲۰۰۵) نیز، در مقالهٔ خود موارد دیگری را یادآور می‌شود، مثل تعداد دفعاتی که هر رکورد را دیگر رکوردهای پایگاه مورد ارجاع قرار می‌دهند و همچنین، به نسبت اصطلاحات مرتبط با تعداد کل اصطلاحات موجود در رکورد اشاره کرده است. (۷: ۴۹)

فرمول دقیقی که طرز کاربرد این معیارها را نشان دهد، الگوریتم (خوارزمی) رتبه‌بندی نامیده می‌شود که از یک موتورجستجو به موتور جستجوی دیگر متفاوت است. به‌دلیل رقابت‌های بین‌المللی در میان شرکت‌های عرضه‌کنندهٔ موتورهای جستجو، برخی از این شرکت‌ها الگوریتم دقیق رتبه‌بندی را افشا نمی‌کنند (۹) و در عوض، الگوریتم‌ها و معیارهایی را که به‌کار می‌برند، به‌گونه‌ای

2. Kim.

کلی‌تر و فارغ از جزئیات در دسترس قرار می‌دهند.

موتور جستجوی هات بوت[3]، بسامد و جایگاه واژگان را به‌عنوان عامل‌های اولیه درنظر می‌گیرد. مدارکی که بسامد بالاتری از واژگان مورد جستجو دارند، وزن بیشتری به خود اختصاص می‌دهند. بسامد واژگان در کل پایگاه هم بر این مسئله تأثیرگذار است. علاوه بر این، تعداد تکرار ویژه نیز نسبت به طول متن درنظر گرفته می‌شود. زمانی که تعداد بسامد واژه برابر باشد، مدارک کوتاه نسبت به مدارک طولانی رتبهٔ بالاتری کسب می‌کنند. همچنین وجود اصطلاحات مورد جستجو در عنوان مدرک و یا در ابربرچسب‌ها[4]، به نسبت زمانی که در متن مدرک وجود دارند، وزن بیشتری به مدارک اختصاص می‌دهند. (۴: ۳۹)

آلتا ویستا[5] نیز، عامل‌های بیان شده را، به نحوی مورد ملاحظه قرار داده و علاوه بر این، تعداد اصطلاحات منطبق شده با واژگان مورد جستجو و همچنین مجاورت واژگان را درنظر داشته است. (۸: ۳۰)

پاره‌ای دیگر از موتورهای جستجو، اطلاعات کمتری در مورد نحوهٔ رتبه‌بندی و عامل‌های آن، ارائه می‌کنند و فقط به تعدادی از عناصر اشاره دارند. از دیدگاه اینفوسیک[6]، وجود اصطلاحات مورد جستجو در عنوان و ابربرچسب ارزش فوق‌العاده‌ای دارد. این، درحالی است که در نگاه لایکوز[7]، اصطلاحات موجود در عنوان و سرعنوان دیده نمی‌شود و اصطلاحات موجود در ابربرچسب، فاقد ارزش است (۴: ۳۹). اکسایت[8] هم، اصطلاحات ابربرچسب را درنظر نمی‌گیرد و علاوه بر بازیابی مدارکی که اصطلاحات جستجو را شامل می‌شود، این موتور جستجو، محتوای مدارک را برای اصطلاحات مرتبط تحلیل می‌کند و این کار، طی فرایندی به‌نام ICE [9] انجام می‌گیرد. (۴: ۳۹)

اخیراً پاره‌ای از موتورهای جستجو رویکردهای دیگری مانند توجه به لینک (پیوند)های رسیده و خارج شده از مدارک مدنظرقرار داده‌اند که از آن میان، موتور جستجوی گوگل را می‌توان نام برد.

طرح تحقیقاتی گوگل

در اوایل سال ۱۹۹۸، دانشجویان دانشگاه استنفورد کالیفرنیا، لارنس پیج و سرجی برین[10]، طی طرحی تحقیقاتی، الگوریتم (خوارزمی) خاصی برای رتبه‌بندی نتایج جستجوها براساس ارتباط بیشتر ، طراحی کردند و آن را در مقاله‌ای با عنوان «کالبدشناسی موتور جستجوی وب فرامتنی در مقیاس بزرگ»[11] در همایش World Wide Web ارائه دادند.

این نوآوری، در ژانویهٔ ۱۹۹۸ به نام لارنس پیج، به شمارهٔ ۶۲۸۵۹۹۹ به ثبت رسید و در

3. HotBot.
4. Meta tags.
5. AltaVista.
6. InfoSeek.
7. Lycos.
8. Excite.
9. Intelligent Concept Extraction.
10. Lawrence Page & Sergey Brin.
11. The anatomy of a large-scale hypertextual; web search engine.

سپتامبر ۲۰۰۱ مورد حمایت مالی قرارگرفت و بدین ترتیب، شرکت گوگل به‌طور رسمی پایه‌گذاری و تأسیس شد. الگوریتم معرفی‌شده، با عنوان PageRank شناخته شد و روزی ۱۵۰ میلیون درخواست را پاسخ می‌گفت. امروزه گوگل با این الگوریتم (خوارزمی)، موفقیت خود را تضمین کرده است. (۵: ۴۲)

الگوریتم (خوارزمی) بازیابی و رتبه‌بندی گوگل

در الگوریتم گوگل، دو بخش اصلی وجود دارد.

بخش اول، سامانهٔ text matching نامیده می‌شود که در آن، موتور گوگل صفحات مرتبط با واژه‌های وارد شده در جعبهٔ جستجو را بازیابی می‌کند.

بخش دوم – که از اهمیت معادلی برخوردار است – سامانهٔ PageRank نامیده می‌شود که طی آن، صفحات ثبت شدهٔ گوگل رتبه بندی می‌گردد. (۳)

سامانهٔ text matching

گوگل در جستجوی اطلاعات مربوط، ارزش و وزن زیادی به برچسب عنوان می‌دهد. این که واژه‌ها یا اصطلاحات مورد جستجو در برچسب عنوان وجود داشته باشد، امر درخور توجهی است. «گوگل، برچسب‌های شبیه کلیدواژه‌ها را درنظر نمی‌گیرد. این، بدان دلیل است که گوگل احتمال می‌دهد که مدیران وبگاه‌ها از این عامل سوءاستفاده کنند و به‌منظور جلب ملاقات‌کنندگان بیشتر، شماری واژه‌های نامربوط در این بخش وارد نمایند» (۲). در مقابل، گوگل، توصیفات و واژگان را از چند خط اول متن روی صفحهٔ وبگاه اخذ می‌کند؛ یعنی اگر کلیدواژه‌های مرتبط در بالای صفحهٔ وب باشند، آن صفحه به عنوان مربوط شناخته می‌شود. (۳)

گوگل، تجمع واژگان مورد جستجو در بدنهٔ اصلی متن و صفحه را نیز، به‌عنوان نشانهٔ ارتباط درنظر می‌گیرد. تجمع ۶ تا ۱۰ درصدی، بهترینِ حالت ممکن است. (۳)

گوگل، اخیراً توجه خاصی به برچسب‌های سرعنوان‌های متعدد ظاهر شده در بخش‌های متن کرده است و به آن‌ها ارزش داده است. ابزار دیگر گوگل برای انتخاب صفحه‌های ذی‌ربط، درنظر گرفتن کلیدواژه‌های مربوط در برچسب‌های برجسته و مشخص است. (۳)

سامانهٔ PageRank

مفهوم PageRank؛ از مراحل آغازین وب جهان‌گستر، هر موتور جستجو، شیوه‌های متنوعی را جهت رتبه‌بندی نتایج جستجو توسعه داده است. درهمین راستا و به‌منظور دستیابی به اهداف رتبه‌بندی مبتنی بر محتوا، موتور گوگل رتبه‌بندی مبتنی بر میزان لینک‌ها را گسترش داد.

برای این منظور، تعداد لینک‌هایی که از متون دیگر به متن حاضر وارد می‌شود، به‌عنوان نقطهٔ اهمیت برای متن و صفحهٔ حاضر به‌شمار می‌آید. هر لینک (پیوند) به صفحهٔ حاضر، به‌عنوان یک رأی مثبت برای آن شمرده می‌شود و هر صفحهٔ وبی – که لینک‌های بیشتری به آن وارد می‌گردد – با

اهمیت‌تر تلقی می‌گردد.

برخلاف مفهومی که از سامانهٔ لینک‌ها استنباط می‌گردد، محاسبهٔ ارزش هر صفحهٔ وب به همین سادگی هم نیست و متغیرهای دیگری هم در آن دخالت دارد. هر صفحه‌ای که به صفحهٔ حاضر لینک (پیوند) داشته باشد، به یک اندازه به صفحهٔ حاضر ارزش نمی‌دهد، بلکه ارزش هر لینک برای سند زمانی مهم است که آن لینک از صفحهٔ باارزشی باشد و آن صفحه در الگوی رتبه‌بندی صفحات، از رتبهٔ بالایی برخوردار باشد. بنابراین، رتبهٔ هر سند به رتبهٔ دیگر اسنادی بستگی دارد که به آن لینک می‌دهند و رتبهٔ آن اسناد هم به همین ترتیب تعیین می‌گردد. این چرخه همچنان ادامه می‌یابد تا اینکه رتبهٔ هر سند به رتبهٔ کل وب بستگی پیدا کند و این، همان ماهیت آزادانهٔ این الگوست. (۱۰)

متغیر دیگری که در این الگوریتم (خوارزمی) دخالت دارد، میزان لینک‌های دیگری است که از صفحه‌ای که به صفحهٔ حاضر لینک دارد، خارج می‌گردد. مثلاً اگر صفحهٔ ما – که قرار است رتبه‌اش محاسبه شود – A فرض گردد و صفحه‌ای که به همین صفحهٔ A لینک داده است، B فرض شود، چنانچه تنها لینکی که از B خارج گردیده، لینک به A باشد، این لینک ارزش بالاتری نسبت به موقعی دارد که از B، ۲۰ لینک خارج گردیده باشد و لینک به A هم یکی از آن ۲۰ لینک بوده باشد.

یعنی هر چقدر صفحه‌ای که به صفحهٔ ما لینک داده یا به اصطلاح رأی داده، لینک‌های بیشتری داشته باشد، رتبهٔ کمتری (PageRank) به صفحهٔ ما تعلق می‌گیرد و برعکس. (۶)

اگرچه این رهیافت به‌نظر خیلی گسترده و پیچیده می‌آید، اما پیج و برین، موفق شدند این را با استفاده از الگوریتم نوآورانهٔ خود عملی سازند.

الگوریتم (خوارزمی) PageRank

فرمول الگوریتم گوگل – که آن را پیج و برین طراحی کردند – به این صورت است:

PR(A) =(1-d) +d (PR(Ti) /C(T) + …..+ PR(Tn) /C (Tn))[۱۲]

PR (A) : PR یا رتبهٔ صفحهٔ A است.

PR (Ti) : PR صفحهٔ Ti که به صفحهٔ A لینک داده است.

C (Ti) : تعداد لینک‌هایی که از صفحهٔ Ti خارج می‌شود.

d : یک damping factor که بین صفر و یک است.

از این فرمول، رتبه‌بندی موارد زیر دریافت می‌گردد:

– این سامانه، صفحات وب را به‌طورکلی رتبه‌بندی نمی‌کند، بلکه رتبهٔ هر صفحه به تنهائی محاسبه می‌گردد و رتبهٔ هر صفحه، به‌وسیلهٔ دیگر صفحاتی که به آن صفحه لینک داده‌اند، محاسبه می‌گردد.

– هرچه تعداد لینک‌های وارد شده به صفحهٔ حاضر بیشتر باشد، رتبهٔ صفحهٔ حاضر بالاتر است.

– هرچه رتبهٔ صفحهٔ لینک داده به صفحهٔ

۱۲. کاربرد PR در مقاله، معادل PageRank است.

حاضر بیشتر باشد، رتبهٔ صفحهٔ حاضر بالاتر می‌رود.

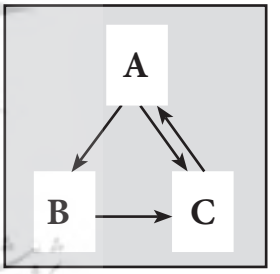
– هرچه تعداد کل لینک‌های صفحهٔ لینک داده به صفحهٔ حاضر بیشتر باشد، رتبهٔ کمتری به صفحهٔ حاضر می‌دهد.

یعنی یک وبگاه با رتبهٔ (PR) ۲ و تعداد ۲ لینک بیش از وبگاهی با رتبهٔ ۶ و تعداد ۱۰ لینک به وبگاه ما ارزش می‌دهد.

مقیاس PageRank (PR)، از یک تا ده است. وبگاه کم اهمیت PR برابر با یک تا سه دارد و وبگاه پراهمیت PR، از هفت تا ده می‌تواند داشته باشد. (۳)

مثال:

یک وب کوچک را، مشتمل بر سه صفحهٔ A، B و C درنظر می‌گیریم:



صفحهٔ A لینک به هر دو صفحهٔ B و C، صفحهٔ B، لینک به صفحهٔ C، و صفحهٔ C، لینک به صفحهٔ A دارد. برطبق نظر پیج و برین، damping factor معمولاً ۰/۸۵ است. اما برای ساده‌تر شدن محاسبه، در اینجا آن را ۰/۵ درنظر می‌گیریم. میزان دقیق damping factor بر PR تأثیرگذار است، ولی بر اصول اساسی آن تأثیر ندارد.

براساس فرمول کلی زیر:

$$PR(A) = (1-d) + d(PR(Ti)/C(T) + \ldots + PR(Tn)/C(Tn))$$

رتبهٔ سه صفحه را محاسبه می‌کنیم:

$$PR(A) = 0.5 + 0.5\ PR(C)$$

$$PR(B) = 0.5 + 0.5\ (PR(A) / 2)$$

$$PR(C) = 0.5 + 0.5\ (PR(A) / 2 + PR(B))$$

بنابراین :

$$PR(A) = 14/13 = 1.07692308$$

$$PR(B) = 10/13 = 0.76923077$$

$$PR(C) = 15/13 = 1.15384615$$

روشن است که جمع PR تمام صفحات، ۳ می‌شود؛ بنابراین معادل با مجموع کل صفحات وب است[۱۳]. تعیین کردن این ارزش برای این سه صفحه راحت بود، ولی درنظر بگیریم که این عملیات در کل وب – که شامل میلیون‌ها سند است – چگونه انجام می‌گیرد؟

موتور جستجوی گوگل، با به خدمت گرفتن شاخه‌های مختلف علم ریاضی از جمله جبر خطی، انجام این عملیات و محاسبات را به بهترین نحو ممکن عملی کرده است.

دیگر عامل‌های تأثیرگذار بر PageRank

از زمان کار علمی سرجی برین و لورنس پیج دربارهٔ PageRank، مباحث زیادی در زمینهٔ عامل‌های دیگری که علاوه بر دستورالعمل‌های لینک‌ها بر PR تأثیرگذار می‌باشد، به‌وجود آمده است. خود لورنس پیج، عامل‌های تأثیرگذار بالقوه‌ای به شرح زیر بیان کرده است:

۱۳. برای اطلاعات بیش‌تر درمورد مثال، به منبع شمارهٔ ۱۰ بنگرید.

۱. لینک‌های قابل مشاهده؛

لینک‌های قابل مشاهده، شامل لینک‌هایی است که برجسته‌ترند و توجه را جلب می‌کنند که از آن میان، به لینک‌های رنگی و یا نوشته شده با قلم ایتالیک، می‌توان اشاره کرد. این لینک‌ها به دلیل اینکه احتمال تلیک کردن تصادفی کاربر روی آن‌ها بیشتر است، از ارزش بیشتری برخوردارند و رتبهٔ بالاتری برای صفحه‌ای که لینک به آن می‌رسد، به‌وجود می‌آورند.

۲. موقعیت لینک‌ها؛

عامل دیگری که می‌تواند بر ارزش لینک تأثیرگذار باشد، موقعیت قرار گرفتن لینک در صفحه است. لینک‌هایی که در نیمهٔ بالای صفحه قرار دارد، نسبت به لینک‌هایی که در نیمهٔ پایین صفحه قرار دارند، از ارزش بیشتری برخوردارند و برعکس. لینک‌های باارزش‌تر، برای صفحهٔ مورد لینک ارزش بیشتری به‌وجود می‌آورند.

۳. فاصلهٔ بین صفحات وب؛

مدیران صفحات وب، ممکن است تأثیرات غیرواقعی در عملکرد PR ایجاد کنند و آن، بدین‌صورت است که مدیران وبگاه‌ها می‌توانند تعداد کثیری از صفحات وب را ایجاد کنند و لینک‌های متقابلی در بین آن‌ها توزیع نمایند که موجب می‌شود رتبهٔ صفحه‌ای خاص به‌طور غیرواقعی در الگوریتم بالا برود. در این حالت، صفحات وب بدون اینکه از صفحات با PR بالا لینک داشته باشند، می‌توانند رتبهٔ بالایی به‌خود اختصاص دهند[۱۴]. در این حالت، نه فقط مفهوم PR تضعیف می‌گردد، بلکه نمایهٔ موتور جستجو نیز با میزان بی‌شماری از صفحات وب بدون کیفیت که فقط برای تأثیرگذاری بر PR به‌وجود آمده‌اند، اشباع می‌گردد. از این‌رو، لورنس پیج در جزئیاتی که در سندش ارائه داده، ارزش خاصی به لینک‌هایی می‌دهد که از صفحاتی با مسافت دورتر می‌رسد و این امر، می‌تواند از تأثیرات غیرواقعی بر PR جلوگیری کند. چون با وجود مسافت بین صفحات، مدیران وبگاه‌ها نمی‌توانند نظارت زیادی بر آن‌ها داشته باشند، یک مقیاس برای مسافت بین دو صفحه می‌تواند براساس Domain باشد. در این حالت، لینک‌های رسیده از داخل Domain، ارزش کمتری نسبت به لینک‌های خارج از آن دارند و لینک‌هایی باارزش‌ترند که از صفحه‌ای خارج از Domain مربوط باشند. برای مسافت، مقیاس‌های دیگری هم می‌توان درنظر گرفت که از آن میان، می‌توان به فاصلهٔ جغرافیایی میان خدمتگرها اشاره کرد.

۴. روز آمد بودن صفحات لینک‌دهنده؛

لورنس پیج، در این‌باره نیز بیان داشته که روزآمد بودن صفحات و اسنادی که لینک را ایجاد کرده‌اند، می‌تواند مبیّن این باشد که اطلاعات مندرج در صفحه – که لینک به آن می‌رسد – تا چه حد روزآمدند و ارزش بیشتری برای صفحه ایجاد می‌کنند.

۱۴ . شاید بتوان این فرایند را با self citation در نوشته‌های چاپی معادل دانست. مؤسسهٔ اطلاعات علمی (ISI) نیز، برای رتبه‌بندی مجلات براساس میزان استناد به مقالات آن‌ها با همین مشکل خود استنادی روبه‌روست و در فرمول جداگانه، موارد خوداستنادی را حذف می‌کند.

الگوی موج‌سوار تصادفی: تأییدی بر الگوریتم (خوارزمی)

لورنس پیج و سرجی برین، تأیید ساده و مستقیمی برای الگوریتمشان ارائه کرده‌اند. آن‌ها PR را به‌عنوان الگویی از رفتار کاربری درنظر گرفته‌اند که با موج سواری در وب، روی لینک‌ها بدون توجه به محتوای آن‌ها تلیک می‌کند. کاربر با موج‌سواری، صفحهٔ خاصی را با یک احتمال مطمئن ملاقات می‌کند که این، از PR صفحه سرچشمه می‌گیرد. احتمال اینکه او به‌طور تصادفی روی یک لینک در یک صفحه تلیک کند، به مدد تعداد لینک‌های روی آن صفحه تعیین می‌گردد. به همین دلیل است که در فرمول PR، هر صفحه به تعداد لینک‌های روی صفحه تقسیم می‌گردد. بنابراین، احتمال اینکه کاربر به‌طور تصادفی به یک صفحه برسد، برابر با جمع احتمالاتی است که در برابر او قرار می‌گیرد تا روی لینک‌ها تلیک کند و به یک صفحهٔ خاص برسد. البته اکنون این احتمال به‌وسیلهٔdamping factor d کاهش یافته است. تأیید برالگوی موج سوار تصادفی[15]، این است که موج سوار در وب روی تعداد مشخصی از لینک‌ها تلیک نمی‌کند، بلکه گاهی خسته می‌شود و به‌طور تصادفی به صفحهٔ دیگری پرش می‌کند.

احتمال برای موج سوار تصادفی با تلیک بر لینک‌های داده شده، به‌وسیلهٔ damping factor d – که بسته به درجهٔ احتمال، بین صفر و یک متغیر است – متوقف نمی‌شود. زمانی که فرد روی لینک‌های تلیک شده باقی می‌ماند، d بالاتر است و زمانی که تصادفی به صفحهٔ دیگر پرش می‌کند، بعد از اتمام لینک‌های تلیک شده، احتمال به‌عنوان یک (1- d) دائم در الگوریتم باقی می‌ماند. همچنین با توجه به لینک‌های واردشده، احتمال پرش تصادفی کاربر به یک صفحه، همیشه (1- d) است. بنابراین، هر صفحه، همواره یک حداقل PageRank را داراست. (۱۰)

ویرایش دوم الگوریتم (خوارزمی) گوگل

دومین ویرایش الگوریتم – که به همت لورنس پیج و سرجی برین منتشر شده – به این قرار است:

$$PR(A) = (1-d)/N + d\,(PR(Ti)/C(T) + \ldots\ldots + PR(Tn)/C(Tn))$$

که در اینجا، N مجموع کامل صفحات وب است . دومین ویرایش، تفاوت بنیادی با ویرایش اول ندارد. دراین ویرایش براساس الگوی موج‌سوار تصادفی، PageRank یا رتبه‌بندی صفحه، یک احتمال واقعی برای موج‌سوار در وب است که براساس آن، او بعد از تلیک کردن روی لینک‌های متعدد به صفحه‌ای خاص می‌رسد. کل PagRrank، یک توزیع احتمال را در کل صفحات وب شکل می‌دهد. بنابراین، جمع کل PageRank صفحات وب، برابر با یک خواهد شد.

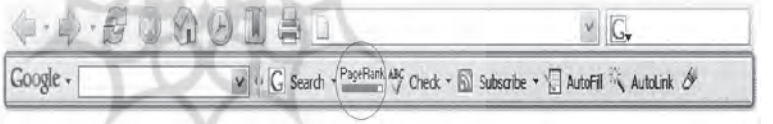
برعکس، در اولین ویرایش الگوریتم، احتمال برای رسیدن تصادفی کاربر به یک صفحهٔ خاص، به کمک جمع تعداد صفحات وب ارزش‌گذاری می‌شود. بنابراین در این

15.The random surfer model.

ویرایش، PR ارزش درخور انتظاری برای موج‌سواری تصادفی کاربری خواهد بود که صفحهٔ خاصی را ملاقات می‌کند و این روند را به هر تعداد صفحه‌ای که وب داشته باشد، دوباره از نو شروع می‌کند. مثلاً اگر وب یکصد صفحه داشته باشد، و یک صفحه با PR برابر با ۲ داشته باشیم، موج‌سواری تصادفی، کاربر را به‌طور متوسط ۲ بار به آن صفحه خواهد رساند، حتی اگر او یکصد بار از نو شروع کند.

نمایش PageRank از نوار ابزارگوگل

نوار ابزار گوگل، به‌عنوان مرورگری است که می‌تواند از وبگاه گوگل دریافت گردد و تعدادی از ویژگی‌ها را برای جستجوی بهتر و راحت‌تر فراهم کند.



این نوار ابزار، PagaRank را روی مقیاس ۱ تا ۱۰ نمایش می‌دهد و این، یکی از خصوصیات پیشرفتهٔ گوگل است. یعنی کاربر در هر صفحه‌ای از وب که حضور داشته باشد، و از طریق موتور جستجوی گوگل آن صفحهٔ خاص را باز کرده باشد، می‌تواند با تلیک کردن بر PagaRank – که بر روی نوار ابزار قرار دارد – رتبهٔ صفحهٔ جاری را مشاهده کند[۱۶].

نتیجه‌گیری

گوگل، بهترین بودن را به‌عنوان نقطهٔ پایان قبول ندارد، بلکه آن را نقطهٔ آغاز می‌داند. اکنون گوگل بر آن است تا دست کم در زمینهٔ موتورهای جستجو پیشرو باشد و برای رسیدن به این هدف، پیوسته نوآوری می‌کند و فن‌آوری خود را گسترش می‌دهد. نمونهٔ بارز آن، الگوریتم (خوارزمی) رتبه‌بندی گوگل است که می‌توانیم بگوییم در عمل بی‌نظیر است و گوگل نیز به آن می‌بالد. این الگوریتم، رتبه‌بندی صفحات را برپایهٔ میزان و کیفیت لینک‌ها قرار داده است و کمیت و کفیت را بر پایهٔ منطق ریاضی محاسبه می‌کند. او این رتبه‌بندی را بر پایهٔ دموکراسی وب می‌داند و تا حدود زیادی هم می‌توانیم بگوییم نسبت به موتور جستجوهای دیگر موفقیت‌های زیادی کسب کرده است. ولی نکتهٔ شایان ذکر در اینجا، این است که با وجود موقعیت کاراً و درخشان این ابزار از لحاظ فنی و عملی، نباید از نظر دور داشت که هنوز هم برای برخی از موضوعات، دقت مضمونی از الگوریتم ریاضی بااهمیت‌تر است.

منابع

۱. پژوهشگاه اطلاعات و مدارک علمی ایران. [قابل دسترسی در]:

۱۶. جهت دان لود مرورگر نوار ابزار گوگل، در صفحهٔ خانگی گوگل، روی گزینهٔ more تلیک می‌کنیم، و بعد از سیاهه‌ای که باز می‌شود، گزینهٔ toolbar را تلیک کنیم و بعد، از منوی بعدی گزینهٔ download را برگزینیم.

7. Kim, Guenther. "Choosing a search engine". ***Online,*** Vol.29, No.1 (Jan./Feb 2005): 48.

8.Ressel, M Robert. "Choosing a search engine for internet exploration". ***The Technology Teacher,*** No.56 (Apr.1997): 30.

9. Sulivon, Danny. "How search engines rank web pages". [on-line]. Available: http://search engine watch.com/web masters/ article. php/ 2167967.

10. "A survey of google pageRank". [on-line]. Available: http://www. misweb design.com/resources/articles/pageRank-8.html.

11. Timothy, Archibald. "Search us, says google". ***Technology Review,*** No.3 (Nov./Dec.2000): 108.

تاریخ دریافت: ۱۳۸۵/۵/۸

http://www.irandoc.ac.ir/ETELAART/ 16/16-3-4-7-htm.

2. Brandt, Daniel. "PageRank: google original sin". [on-line]. Available: http:// www.google-watch.org/pagerank.html.

3. Callan, David. "Google ranking tips". [on-line]. Available: http:// www. Akamarketing .com/google-ranking.

4. Courtois, Martin P; Berry, Michael W. "Result ranking in web search engines". ***Online***, Vol.23, No.3 (May/Jun.1999): 39.

5. Demetios, Eliopoulos; Gotlieb, Calvin. "Evaluating web search ranking". ***Online,*** Vol.27, No.2 (Mar./Apr. 2003): 42.

6. Google technology. [on-line]. Available: http://www.google.com/ technology/index.html